

Geographic knowledge extraction and semantic similarity in OpenStreetMap

Andrea Ballatore · Michela Bertolotto ·
David C. Wilson

Received: 17 October 2011 / Revised: 10 August 2012 / Accepted: 6 October 2012 /
Published online: 23 October 2012
© Springer-Verlag London 2012

Abstract In recent years, a web phenomenon known as Volunteered Geographic Information (VGI) has produced large crowdsourced geographic data sets. OpenStreetMap (OSM), the leading VGI project, aims at building an open-content world map through user contributions. OSM semantics consists of a set of properties (called ‘tags’) describing geographic classes, whose usage is defined by project contributors on a dedicated Wiki website. Because of its simple and open semantic structure, the OSM approach often results in noisy and ambiguous data, limiting its usability for analysis in information retrieval, recommender systems and data mining. Devising a mechanism for computing the semantic similarity of the OSM geographic classes can help alleviate this semantic gap. The contribution of this paper is twofold. It consists of (1) the development of the OSM Semantic Network by means of a web crawler tailored to the OSM Wiki website; this semantic network can be used to compute semantic similarity through co-citation measures, providing a novel semantic tool for OSM and GIS communities; (2) a study of the cognitive plausibility (i.e. the ability to replicate human judgement) of co-citation algorithms when applied to the computation of semantic similarity of geographic concepts. Empirical evidence supports the usage of co-citation algorithms—SimRank showing the highest plausibility—to compute concept similarity in a crowdsourced semantic network.

Keywords Semantic similarity · OpenStreetMap · Volunteered Geographic Information · OSM Semantic Network · SimRank · P-Rank · Co-citation · Crowdsourcing

A. Ballatore (✉) · M. Bertolotto
School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
e-mail: andrea.ballatore@ucd.ie

D. C. Wilson
Department of Software and Information Systems, University of North Carolina,
University City Boulevard, Charlotte, NC, USA

1 Introduction

For almost a decade, digital geographic information has experienced enormous expansion reaching millions of Internet users, well beyond the limited circles of geographers, cartographers and urban planners, who have traditionally been the gatekeepers. This rapid growth of broader interest and engagement in geographic information has been studied from several viewpoints: crowdsourced data collection or Volunteered Geographic Information (VGI), ubiquitous cartography, web mapping, and wikification are all facets of this complex phenomenon that Turner has named *neogeography* [59].

In the mutable constellation of neogeography, OpenStreetMap (OSM) has emerged as the most ambitious and, in some respects, the most successful collaborative online project [21]. Through a Wiki model adapted for spatial data, its numerous users create, edit and utilise a vector map covering the entire planet. Although the project name suggests an emphasis on routing, the map includes natural entities and man-made features, from the borders of nations to the post boxes in rural towns.

Given this extremely wide scope, it is clear that one of the critical aspects for the coherence and quality of the OSM vector data lies in its associated semantic structure. In the OSM vector data set, map objects are associated with properties that encode their semantic content, structured in key/value pairs (e.g. *amenity=university*, *name='University College Dublin'*). In OSM, the properties of an object are called 'tags'.¹ The meaning and usage of these tags are negotiated within the contributors' community on the OSM Wiki website.² The meaning of the tags can change over time, in a process of *emergent semantics*, where concepts emerge, shift and disappear in a complex evolutionary negotiation [38].

With the emergence of the Semantic Geospatial Web, the computation of semantic similarity has gained prominence in the Geographical Information Science (GIScience) community [12]. Semantic similarity has attracted remarkable interest within several academic disciplines, originally in psychology, and subsequently in linguistics, cognitive science and knowledge engineering [27,28]. Several measures tailored to geographic concepts have been proposed and evaluated [15,50,52].

An effective measure of semantic similarity between OSM geoconcepts can facilitate the usage of OSM data in numerous applications, such as geographic information retrieval, spatial recommender systems, data mining, location-based services and geo-information integration. For example, given three classes of entities that can commonly be found in maps, *fountain*, *school* and *bookshop*, a semantic similarity measure is expected to find a stronger association between *school* and *bookshop* than between these two concepts and *fountain*. However, only considering the current OSM tag structure, this distinction cannot be captured because all of these three concepts are siblings under the same parent concept (*amenity*).

Some projects address the need for semantic support in OSM. LinkedGeoData has republished the OSM map as a Semantic Web data set, structured on a shallow tree representing the tags [5], while OSMonto³ consists of a formal description of a subset of OSM tags. None of these projects exploits the OSM Wiki website as a source of semantic knowledge.

Indeed, the OSM Wiki website contains a densely connected graph of pages describing geographic concepts. While semantic information is *implicitly* present in the Wiki link

¹ This usage of the term 'tag' is highly unusual: in the Web 2.0, tags are generally unstructured text labels used as meta-data [22]. However, to be consistent with the OSM terminology, we will refer to the OSM properties as 'tags' in the rest of this paper.

² <http://wiki.openstreetmap.org> (acc. August 10, 2012).

³ <http://wiki.openstreetmap.org/wiki/OSMonto> (acc. August 10, 2012).

structure, it has to be made *explicit* in order to exploit it and to refine the computation of semantic similarity. In this framework, co-citation algorithms [29,54] seem promising to compute semantic similarity in a semantic network, but have been neglected in favour of other measures, particularly in the geospatial field [52]. To the best of our knowledge, no in-depth study on their cognitive plausibility has been published. In order to address these points, our contribution to the area of VGI semantics and semantic similarity consists of two parts:

1. The development of the OSM Semantic Network,⁴ by means of a dedicated, open source web crawler. This network captures semantic relationships between geographic concepts, which are implicit in the OSM Wiki website. Among other applications, it allows the measurement of the semantic similarity between concepts.⁵ It, therefore, represents a useful support tool for geographic information retrieval, recommender systems and data mining.
2. A study on cognitive plausibility (i.e. the ability to mimic human behaviour) of co-citation algorithms to compute semantic similarity of geospatial classes. While this approach can be in principle applied to any network, our experiments have been conducted on the OSM Semantic Network.

The remainder of this article is organised as follows: Sect. 2 reviews related work in the areas of VGI, OSM, and semantic similarity measures. Section 3 discusses OSM semantics, while Sect. 4 presents the OSM Semantic Network. Section 5 frames the idea of co-citation in the context of OSM, and Sect. 6 presents the study of cognitive plausibility of co-citation algorithms. Finally, Sect. 7 draws conclusions from this work and suggests directions for future research.

2 Related work

Our research is positioned at the intersection between VGI, OSM, and the existing approaches to semantic similarity, particularly within the area of GIScience. This section provides an overview of these areas, highlighting related work.

2.1 Volunteered Geographic Information (VGI)

During the past decade, the rapid expansion of Web 2.0 has resulted in several crowdsourcing phenomena, such as folksonomies, wiki models, social tagging, social bookmarking and collaborative classification [22,49]. Digital geographic information has also experienced unprecedented growth, both in quantitative and qualitative terms. Goodchild [18] termed the crowdsourcing of geographic information as Volunteered Geographic Information, specifically referring to geographic information produced and released by non-expert users through *voluntary* actions.

VGI is having a visible impact on the production and consumption of geographic information, adding a collaborative dimension to the traditionally hierarchical, centralised model of production [20]. In parallel, the expansion of mapping to increasingly powerful mobile computing devices has led to the so-called *ubiquitous cartography* [16].

⁴ <http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork> (acc. August 10, 2012).

⁵ Pre-computed similarity scores for the entire OSM Semantic Network are available at <http://spatial.ucd.ie/osn/similarities> (acc. August 10, 2012).

Discussing these trends, Sui [57] suggests the term *wikification* to capture the attempt to crowdsource non-textual data, emulating Wikipedia within the spatial domain. Friedhorsky and Terveen proposed an adaptation of the wiki model for spatial data [47]. The growth of available online geographic information raises the issue of semantics: data are useless unless its meaning is intelligible. The threat of a deluge of semantically poor geo-data prompted Egenhofer [12] to envisage the emergence of the Semantic Geospatial Web, a spatial extension of the Semantic Web initiative that will enable advanced information retrieval. *Neogeography* is the umbrella term that Turner [59] coined in order to discuss this nexus of phenomena.

Overall, neogeography can be defined as crowdsourced, wikified, interactive, web-based, volunteered and ubiquitous. Several neogeographic online projects gathered wide communities of users and contributors. These projects range from the very specific—Cyclopath collects cycling-related geographic knowledge—to the very generic—Wikimapia is a commercial effort to build an online editable map where the user ‘can describe any place on Earth’.⁶ Among these initiatives, we focus on OpenStreetMap, the only large-scale attempt at creating a fully open-content vector world map.

2.2 OpenStreetMap (OSM)

Arguably, the leading VGI initiative OSM aims at creating an open vector map of the world [21]. Unlike other VGI projects, OSM revolves around the construction of a vector data set representing the entire planet, not just annotations on an existing map, and emphasises the openness of its datasets.⁷

Since its inception in 2004, OSM has been growing at a considerable rate, attracting attention both in academia and in industry [46]. Given the project’s reliance on the Wiki model, one of the most significant issues is data quality, which for the moment remains an open problem [19,43]. While the geometric quality of OSM data is debated, little work has been done on the *semantic* quality of the classes under which the geometries are classified. The OSM geometries are described through tags, which indicate their meaning and functional role in the data set. For example, a university campus consists of a polygon delimiting its boundaries, associated with the tag *amenity=university* (see Sect. 3).

A set of similar and sometimes competing projects has emerged to enhance OSM semantics. LinkedGeoData (LGD) has taken the entire OSM data set and republished it in a Semantic Web-friendly format, linking it to a formal ontology.⁸ Despite the advantages of the new format, the LGD ontology is a simple, shallow tree structure, representing keys and values. Its semantic content is limited to *is_a* relationships between tags and respective values. OSMonto⁹ offers another ontology based on OSM tags. Its main data set consists of an incomplete formal description of a subset of OSM tags.

To date, none of these projects has been officially integrated into the OSM infrastructure, and OSM semantics has been largely left unexplored. Furthermore, to the best of our knowledge, none of the aforementioned projects provides a semantic similarity measure for OSM geographic concepts.

⁶ <http://cyclopath.org>, <http://wikimapia.org> (acc. August 10, 2012).

⁷ http://wiki.openstreetmap.org/wiki/OpenStreetMap_License (acc. August 10, 2012).

⁸ <http://linkedgeo.org/ontology> (acc. August 10, 2012).

⁹ <http://wiki.openstreetmap.org/wiki/OSMonto> (acc. August 10, 2012).

2.3 Similarity measures

Similarity is a ubiquitous concept in computing. Clustering, information retrieval, pattern recognition, data mining, image analysis and recommender systems rely heavily on some measure of similarity between text documents, images, vectors, concepts and other digital objects [30,36,53,60]. Wittgenstein remarked that the meaning of words flows through “a complicated network of similarities overlapping and criss-crossing,” rejecting the idea that concepts can be given clear and definitive boundaries [61, Par. 66]. Today, it is quite uncontroversial that such a ‘complicated network of similarities’ occupies a central position in human cognition and thinking [52].

In the field of GIScience, semantic similarity measures enable data integration from different sources, ontology alignment, data mining and semantic information retrieval, that is, dealing with ambiguous and fuzzy queries [11,15,28]. Schwering [52] has proposed a classification of semantic similarity approaches. In her view, *feature models* interpret objects as unstructured sets of features and compute their similarity on set-theoretical measures. The Matching-Distance Similarity Model (MDSM) extends the ratio model developed by Tversky taking into account different features of a geographic concept (parts, functions and attributes) [50]. Moreover, Janowicz et al. [26] have developed SIM-DL, a feature model based on description logics.

While approaches such as MDSM and SIM-DL measure similarity on the ontological description of geographic classes, semantic similarity cannot be assessed at the instance level. Mülligann et al. [44] extract a similarity measure directly from the OSM vector data set, looking at the spatial co-occurrence of features. However, the scope of their study is restricted to points of interest, and there is no evaluation on how this measure correlates with human judgement.

Network models are used to measure similarity in semantic networks. Semantic networks encode knowledge and meaning in the form of graphs, whose vertices represent concepts [55]. Such models have been widely used in psychology and cognitive science, for example, to study the workings of human semantic memory [9]. These approaches to similarity are based on some form of structural distance between nodes (e.g. edge counting), sometimes adding additional parameters to weigh the paths [48], or on the topological comparison of subgraphs [35].

Such network-based techniques generally rely on well-defined, expert-generated semantic networks such as WordNet [39]. However, many real-world data sets on the Internet do not present such a structure, but encode valuable information in the form of graphs of inter-linked objects, sometimes referred to as *information networks*. Given the spread of such networks in many fields, algorithms have emerged to identify similar objects exclusively on their link patterns in a network that does not explicitly encode attributes, parts and other details of concepts.

2.4 Co-citation algorithms

In 1973, Small published the *co-citation* algorithm [54]. Given a directed graph representing scientific papers and their mutual citations, co-citation measures the similarity between two given papers by the frequency in which they are cited together. Extending co-citation to an iterative form, Jeh and Widom [29] in 2002 created *SimRank*, a structural approach to calculate vertex similarity in directed graphs. The underlying recursive assumption is that “two objects are similar if they are referenced by similar objects” [29, p. 541]. Given its generality and effectiveness, SimRank has attracted notable research interest [33,37].

The *P-Rank* algorithm (Penetrating Rank) generalises SimRank, taking into account outgoing links, stating that “two entities are similar if (1) they are referenced by similar entities; and (2) they reference similar entities” [62, p. 553]. Classic algorithms such as the original Co-citation [54], Coupling [31], and Amsler [4] are specific cases of P-Rank. As recent surveys within GIScience do not address these approaches, the community does not seem to have explored their potential to assess semantic similarity within the geographic domain, favouring other models [27, 52].

When computing the semantic similarity, it is essential to assess how a computational measure correlates with human thinking (i.e. cognitive plausibility). The cognitive plausibility of semantic similarity measures for geo-concepts has been studied for MDSM [50] and SIM-DL [25]. To the best of our knowledge, the cognitive plausibility of co-citation algorithms applied to semantic networks has not been investigated.

3 OSM Semantic Network extraction

This section describes the development of a new semantic network by means of a dedicated web crawler, tailored to the OSM Wiki website. In the OpenStreetMap vector data set, map objects are encoded as *nodes* (points of interest or centroids), *ways* (lines and polygons) and *relations* (groups of objects). The world data set currently contains 1.2 billion nodes, 106 million ways and 1 million relations.¹⁰ Every map object is described through properties called ‘tags’, defining the semantic content of the object (e.g. *amenity=university*).

The OSM tags are proposed, defined, discussed and sometimes discarded on the OSM Wiki website, which hosts detailed definitions and usage guidelines.¹¹ This website is used as a reference to document and facilitate the mapping process, which is conducted through separate, dedicated web services and tools, which are outside the scope of this paper. According to the OSM Wiki website, tagging should deliberately be informal, loose and open. Mappers are encouraged to use well-known tags, but they are not discouraged from creating new tags when it is deemed useful. This is a more radical policy than that of comparable projects, such as Wikimapia (see Sect. 2).

The OSM keys can represent groups of geographic entities (e.g. *waterway*, *landuse*, *natural*) or encode properties with unrestricted values (e.g. *name*, *addr:street*). While some keys have a small set of well-defined values (e.g. *junction*), other keys have become very large, overstressing their semantic boundaries. The key *amenity*, for example, is associated with more than 150 values, ranging from fast food restaurants to hospitals and cinemas. Moreover, similar tags can be defined with different keys, resulting in semantic difficulties for the users (e.g. *landuse=garages* versus *amenity=parking*). This semantic gap, occasionally, can cause disagreements among users, resulting in ‘tag wars’ [42].

To date, the OSM community has about 453,000 contributors. Through the OSM Wiki website, this large group negotiates what Kuhn [32] calls the ‘social agreements’ needed to define common semantic symbols that can be understood by most users. The fluid openness of OSM semantics is both a strength and a weakness of the project. While contributors are attracted to the lack of formal validation procedures to make changes to the map, this degree of freedom generates noise in the form of semantic ambiguity and redundancy. For this reason,

¹⁰ <http://wiki.openstreetmap.org/wiki/Statistics> (acc. August 10, 2012).

¹¹ http://wiki.openstreetmap.org/wiki/Map_Features (acc. August 10, 2012).

several efforts have been undertaken to monitor the tag usage in the vector data set, such as the web services TagInfo and TagWatch.¹²

The OSM Wiki website encodes semantic content as a collection of inter-linked pages, discussing aspects of the OSM vector data set. Textual descriptions, images and links to Wikipedia are used by contributors to clarify the meaning and usage of OSM tags. In this sense, the OSM Wiki can be seen as a semantic network, in which the pages are concepts and the links represent relationships [55]. In such a network, concepts have connections with other concepts. As pages are modified and reconnected to other pages, the network topology changes accordingly. In the development of this semantic network, we focused on *key* and *tag pages* in English. The OSM Wiki pages can be categorised as follows:¹³

1. **Key page.** Describes the meaning and usage of an OSM key, grouping several tags with the same key. For example, the page `osmwiki:Key:amenity` summarises the key *amenity* and its recommended values (e.g. *university*, *pub*).
2. **Tag page.** Describes a specific key/value pair, representing a concept in the semantic network. For example, `osmwiki:Tag:amenity=library` defines the tag *amenity=library*.
3. **Proposed tag page.** Some tags have been proposed by contributors and are undergoing review. For instance, the tag *historic=aqueduct* has been proposed in `osmwiki:-Proposed_features/aqueduct` and is currently marked as a draft.
4. **Cluster pages.** Pages that group related links to tag pages, while not representing directly a tag (e.g. `osmwiki:Building_attributes`).
5. **Other pages.** All the pages that do not fall in the previous categories, including contributor profiles, technical pages unrelated to tags, and administrative pages (e.g. `osmwiki:Linear_maps`).

More formally, the OSM Wiki can be conceptualised as a directed graph $G = (V, E)$, where vertices V are the web pages, and edges E are their hyperlinks. In order to extract the directed graph G from the OSM Wiki, we implemented the OSM Wiki Crawler, an open source tool tailored to the OSM Wiki content structure.¹⁴

3.1 The OSM Wiki crawler

The purpose of this semantic crawler is the extraction of a semantic network from a dynamic and complex wiki website, encoding geographic knowledge that can be utilised for various tasks—in this paper, we focus on the computation of semantic similarity. Although the crawler focuses on the OSM Wiki website, its general approach can be adopted to extract a semantic network from any wiki, open-content websites.

The extracted network is stored in the resource description framework (RDF), containing a set of statements of the format *(subject, predicate, object)*, logically equivalent to a labelled, directed graph.¹⁵ The crawler downloads and analyses the XML dump provided by OSM, which contains the complete content of the website.¹⁶ To date, the OSM Wiki website is made of about 30,000 pages, 5,500 of which describe key and tags used in the vector

¹² <http://taginfo.openstreetmap.org>, <http://wiki.openstreetmap.org/wiki/Tagwatch> (acc. August 10, 2012).

¹³ ‘osmwiki:’ stands for the namespace <http://wiki.openstreetmap.org/wiki/> (acc. August 10, 2012).

¹⁴ <http://github.com/ucd-spatial/OsmWikiCrawler> (acc. August 10, 2012).

¹⁵ <http://www.w3.org/RDF> (acc. August 10, 2012).

¹⁶ <http://dump.wiki.openstreetmap.org> (acc. August 10, 2012).

map. The crawler extracts from each page the following information, if available: OSM keys and tags, lexical descriptions, relationships between tags, general internal links, and links to Wikipedia pages.

A heuristic function assigns OSM tags to the equivalent terms in the LinkedGeoData light-weight ontology [5]. The heuristic is based on lexical matching between the OSM tag and the LinkedGeoData term. For example, the OSM tag *amenity=fountain* is matched against *lgdo:AmenityFountain*.¹⁷ If the *key=value* pair is not defined, only the value is considered (e.g. *lgdo:Fountain*). We have validated this approach by observing that, in a random sample of size 30, all the mappings to LinkedGeoData were correct.¹⁸

4 OSM Semantic Network

The open source tool that we have developed, the OSM Wiki Crawler, extracts a semantic network from the OSM Wiki website, in the form of an RDF graph. The graph vertices represent OSM keys, tags and clusters. The edge labels specify a number of different relationships between vertices, ranging from links to a tag key (*osmwiki:key*) to a logical implication (*osmwiki:implies*). Generic internal hyperlinks (*osmwiki:link*) are particularly important, as they capture general relatedness between the source and the target pages, useful to compute a cognitively plausible semantic similarity. For example, *amenity=library* contains generic links to *tourism=museum* and *shop=books*.

Cluster pages do not represent tags directly, but contribute to the computation of the semantic similarity between tags. For instance, the cluster page on building attributes¹⁹ strengthens the connectivity between several tags related to buildings. To promote semantic interoperability, the OSM Semantic Network is designed as Linked Data.²⁰ OSM terms are linked to Wikipedia pages and LinkedGeoData terms to which they are semantically equivalent (e.g. *osmwiki:Tag:amenity=embassy* is linked to <http://en.wikipedia.org/wiki/Embassy> and to *lgdo:Embassy*). The detailed content of the RDF graph is presented in Table 1. Pre-extracted networks are available online.²¹

In order to extract information from the semantic network, it is useful to look at its statistical properties. Defining the degree of a vertex as the number of incident edges (formally $d_G(v)$), the mean degree in G is 9.66, which indicates that OSM tags are strongly interconnected. Furthermore, the indegree and outdegree of a vertex can be defined as the number of its ingoing and outgoing edges. In the OSM Semantic Network, the mean indegree is 3.6, while the mean outdegree is 6.06. Figure 1 shows the degree distribution in the OSM Semantic Network, divided into (a) indegree and (b) outdegree. The figure shows that both quantities roughly follow a power law distribution.

This is consistent with the results reported by Broder et al. [8] in 2000: Representing the entire World Wide Web as a directed graph, the degree distribution follows a power law, that is, most pages have low connectivity, while few pages have high connectivity. Interestingly, the OSM Wiki also shares this characteristic with Wikipedia, whose degree distribution also follows a power law, in particular, the Zipf distribution [45]. By treating Wikipedia as a semantic network, it is possible to measure relatedness between pages [45] or to find missing

¹⁷ ‘lgdo:’ stands for the namespace <http://linkedgeo.org/ontology/> (acc. August 10, 2012).

¹⁸ The full algorithm of the crawler is available in the source code documentation.

¹⁹ *osmwiki:Proposed_features/Building_attributes* (acc. August 10, 2012).

²⁰ <http://linkeddata.org> (acc. August 10, 2012).

²¹ <http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork> (acc. August 10, 2012).

Table 1 OSM Semantic Network vertices (total: 6,467) and edges, sorted by number of instances (total: 28,807)

URI	Description	Instances
<i>Vertices</i>		
osmwiki:Key:{key}	OSM key	1,503
osmwiki:Tag:{key = value}	OSM tag	2,047
osmwiki:Proposed_features/{tag}	OSM proposed tag	784
osmwiki:{page}	OSM cluster page	22
others	LGD and Wikipedia nodes*	2,111
<i>Edges</i>		
osmwiki:link	Internal hyperlink within OSM Wiki	12,974
osmwiki:key	Link to OSM key page	5,408
rdf:rdf-schema#comment	OSM tag description	2,889
osmwiki:combinedWith	Tag is combined with target tag	2,054
osmwiki:wikipediaLink	A link to a Wikipedia page	1,604
owl:owl#equivalentClass	Equivalent class in other ontology	652
osmwiki:implies	Tag implies target tag	226

Vertices marked with * are leaf vertices, that is, have only incoming edges. Graph extracted on 1 February 2012

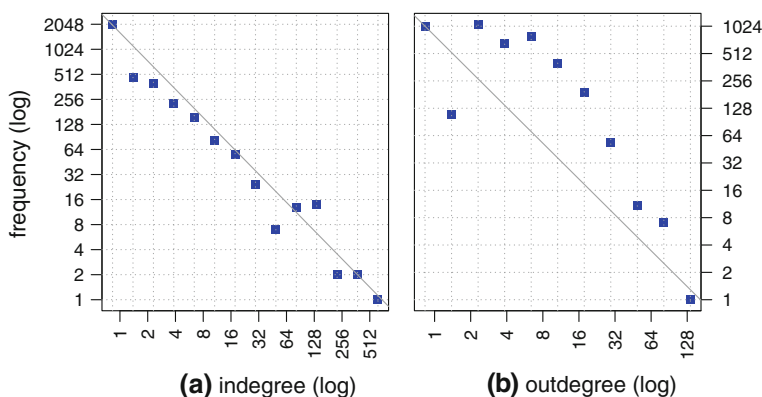


Fig. 1 Distribution of vertex degree in the OSM Semantic Network considering (a) incoming edges and (b) outgoing edges. As shown graphically, these distributions can be approximated by a power law $p(x) = bx^{-\alpha}$. Graph extracted on 1 February 2012

links [1]. Turdakov and Velikhov [58] use the Dice measure on Wikipedia hyperlinks to retrieve semantically related articles and also to perform word sense disambiguation. It is, therefore, not unreasonable to expect that such a dense link structure in the OSM Semantic Network contains information about the semantic similarity between OSM tags.

5 Co-citation for OSM tag similarity

The OSM Semantic Network represents tags and their mutual connections. To exploit the semantic content of the network, we explore its potential to compute the semanticsimilarity

of OSM tags. We define a similarity measure between the tags a and b as $s(a, b) \in [0, 1]$, where 0 means no similarity, and 1 means maximum similarity being a and b vertices in the OSM Semantic Network. In this article, we focus on a tag-to-tag similarity measure, leaving the object-to-object case for future work.

Network-based similarity techniques assume that the relationships between concepts must be sufficiently rich and representative [52]. To assess whether its dense link structure contains valid knowledge about the OSM tags, we compute a similarity score purely based on the network topology, ignoring the lexical descriptions of tags. Approaches such as MDSM and SIM-DL (see Sect. 2.3) have been devised specifically for geographic concepts. Because such measures require a detailed description of attributes, parts and roles not present in OSM, they cannot be used in this context.

Because of the shallow semantic structure of OSM, visible both in LinkedGeoData and OSMonto, the paths between OSM tags are very short: The majority of concepts are connected through 2 edges, even when semantically very dissimilar (e.g. *sauna* is linked to *amenity*, which links back to 150 values, including *bench*). Shortest path-based techniques need paths of variable lengths to be effective and are, therefore, doomed to fail in this case. To compute the semantic similarity of OSM tags, it is necessary to identify alternative measures. Co-citation-based algorithms seem promising.

5.1 Co-citation in a semantic network

As shown in Sect. 2.3, co-citation algorithms aim at finding similarity in a graph of inter-linked objects, based on the intuition that similar objects are referenced together. Although it is possible to compute co-citation measures on the LinkedGeoData and OSMonto ontologies, this would result in a binary classification between tags that are in the same subtree (e.g. *amenity=school*, *amenity=fountain*) or not (e.g. *amenity=school*, *landuse=forest*). This approach is unable to account for semantic similarity within the same key, for example, *amenity=school* and *amenity=university* are expected to be more similar than *amenity=school* and *amenity=fountain*. On the other hand, our OSM Semantic Network allows for a finer computation of similarity by including general hyperlinks between pages and can distinguish between these cases.

Co-citation algorithms have not been utilised to compute semantic similarity of geographic classes. To fill this knowledge gap, we consider *P-Rank*, a generic co-citation algorithm [62]. By setting different values to its parameters, P-Rank is equivalent to earlier algorithms, including Co-citation [54], Coupling [31], and Amsler [4], SimRank [29], and rvs-SimRank [62]. For this reason, it is possible to observe the performance of co-citation algorithms by exploring the result space of P-Rank. In this context, we propose a linear algebra formulation of P-Rank, discussing in detail the meaning and impact of its parameters (K , λ , and C), largely left implicit in the literature [29, 33, 62].

P-Rank is a recursive measure of similarity, based on the combination of two recursive assumptions: (1) two entities are similar if they are referenced by similar entities; (2) two entities are similar if they reference similar entities. All of the notations and symbols used in this paper are summarised in Table 2.

P-Rank is calculated iteratively, choosing a number of iterations $K \in [1, \infty)$. The higher the K , the better the approximation of the theoretical solution to P-Rank. At the first iteration R_0 ($k = 0$), the scores are initialised to 0, $R_0(a, b) = 0$, apart from the identities (if $a = b$, then $R_0(a, b) = 1$). All P-Rank iterations with $k > 0$ can be expressed as a series of iterations converging to the theoretical similarity score:

Table 2 Notations

Symbol	Description
$\mathbf{G} = (V, E)$	The directed graph in which each vertex $a \in V$ represents a OSM tag and $\langle a, b \rangle \in E$ is a hyperlink from tag a to b
$s(a, b)$	Similarity score between tags a and $b \in V$. $s(a, b) \in [0, 1]$, $s(a, b) = s(b, a)$. When $a = b$, $s(a, b) = 1$
$I(a)$	Set of incoming links to tag $a \in V$. $ I(a) $ is the indegree of a
$O(a)$	Set of outgoing links to tag $a \in V$. $ O(a) $ is the outdegree of a
C	P-Rank decay factor. $C \in (0, 1)$. If $C = 1$, P-Rank does not converge
λ	P-Rank in–out balance constant. $\lambda \in [0, 1]$. $\lambda = 1$: incoming links; $\lambda = 0$: outgoing links
k	P-Rank current iteration. $k \in [0, K]$
K	P-Rank maximum iterations. $K \in [1, \infty)$
\mathbf{R}_k	P-Rank score matrix at iteration k
\mathbf{T}_i	Transition matrix of \mathbf{G} constructed on $I(a)$
\mathbf{T}_o	Transition matrix of \mathbf{G} constructed on $O(a)$
Θ	Diagonal matrix. $\forall k$, when $a = b$, $\Theta(a, b) + \mathbf{R}_k(a, b) = 1$

$$s(a, b) = \lim_{k \rightarrow \infty} \mathbf{R}_k(a, b) \quad (1)$$

$$\mathbf{R}_k = C(\lambda \cdot \mathbf{T}_i \mathbf{R}_{k-1} \mathbf{T}'_i + (1 - \lambda) \cdot \mathbf{T}_o \mathbf{R}_{k-1} \mathbf{T}'_o) + \Theta$$

The similarity $s(a, b)$ is a function $f(C, \lambda)$. The constant C is the decay factor applied to the recursive propagation of similarity across the edges. When C is close to 0, almost no similarity flows from one pair to its neighbours, while with C close to 1, the opposite situation arises. The constant λ , on the other hand, is the in–outlinks balance. When $\lambda = 1$, only incoming links are considered, while $\lambda = 0$ indicates that the similarity is computed only on the outgoing links. The number of iterations K determines the minimum precision of the algorithm, that is, the maximum gap between $s(a, b)$ and $\mathbf{R}_k(a, b)$, which decreases as K grows [37]. K , while obviously influencing $\mathbf{R}_k(a, b)$, has no impact on $s(a, b)$.

6 Cognitive plausibility of co-citation algorithms

In this section, we describe an experimental study on the cognitive plausibility of co-citation algorithms, in the case of the computation of semantic similarity of geographic classes. Following Janowicz et al. [25], we define a quantitative measure of *cognitive plausibility* as the observable correlation between the machine-generated rankings of concept pairs and human-generated rankings, ignoring the underlying mental operations.

This approach to cognitive plausibility was originally developed in the area of computational linguistics: several sets of word pairs ranked by humans have been published as ‘gold standards’ against which the similarity measures can be tested. Rubenstein and Goodenough [51] have collected a set of 65 word pairs ranked by their synonymity; Miller and Charles [40] published a similar dataset with 30 word pairs. The WordSimilarity-353 dataset contains 353 word pairs, ranked by similarity and relatedness [2]. However, none of these data sets fits our context, as they contain few words related to geographic entities.

In GIScience, similarity data sets have been created assessing the similarity of geographic concepts. In their evaluation of the SIM-DL algorithm, Janowicz, Keßler, Panov, Wilkes, Espeter and Schwarz [25] have collected human rankings for concepts related to bodies of

water. This data set would not fit our evaluation because it is restricted to a specific geographic semantic subdomain (bodies of water), and it was collected through a questionnaire in German—in this paper, we consider only OSM semantics in English.

6.1 MDSM evaluation data set

This geographic similarity data set, originally collected by Rodríguez and Egenhofer, is suitable to study the cognitive plausibility of co-citation measures. The data set was utilised to evaluate MDSM, their semantic similarity measure [50]. They collected similarity judgements for 33 geographic concepts, including large natural entities (e.g. *mountain* and *forest*), and man-made features (e.g. *bridge* and *house*). Because these concepts were defined in an abstract way through a short lexical definition (without focusing on ontology-specific information), they are suitable to study the cognitive plausibility of our approach.

Judgements were obtained from 72 students through two surveys (*A* and *B*), each presenting five questions. Each question consists of a target concept (e.g. *stadium*) and 10 or 11 base concepts to sort according to their similarity to the target. The results indicate the ranking of the concept pairs, from the most to the least similar (e.g. $\langle \textit{athletic field}, \textit{ball park} \rangle \rightarrow \dots \rightarrow \langle \textit{athletic field}, \textit{library} \rangle$). In their evaluation, Rodríguez and Egenhofer focused the impact of context on similarity judgment. We excluded from the MDSM data set four questions that specify a particular context, which is beyond the scope of this paper.

The 33 concepts of the MDSM data set were manually mapped onto the corresponding tags in the OSM Semantic Network, based on their textual definitions. For example, the concept *tennis court* was matched to `osmwiki:Tag:sport=tennis`. While 29 concepts have a satisfactory equivalent in OSM, four concepts (*terminal*, *transportation*, *lagoon* and *desert*) were discarded because they did not have a precise matching concept in the OSM Semantic Network. As a result, we obtained a modified MDSM data set containing five questions on 29 geographic concepts. The entire data set is available online, including the complete manual mapping and definitions.²²

6.2 Experiment setup

To obtain semantic similarity scores for the OSM tags, we have run several co-citation algorithms on the OSM Semantic Network described in Sect. 4. P-Rank [62] is a generic algorithm that, with certain combinations of parameters C , K , and λ , is equivalent to Co-citation [54], Coupling [31], Amsler [4], SimRank [29], and rvs-SimRank [62]. Hence, in order to study the cognitive plausibility of these algorithms, we have computed P-Rank for 550 unique combinations of K , C and λ . The experimental setup is the following (see Table 2 for notations):

- λ : 11 discrete equidistant levels $\in [0, 1]$.
- C : 5 discrete equidistant levels $\in [.1, .9]$.
- K : 10 P-Rank iterations.

Following the approach adopted by Rodríguez and Egenhofer [50], the results were computed on the rankings and not on the similarity scores, that is, the order of the pairs returned by the system against the order in the MDSM data set, using Spearman's rank correlation coefficient [56]. Spearman's ρ was computed on each of the five questions, over the 550 combinations. To assess how the algorithms performed overall, a meta-analysis of correla-

²² <http://github.com/ucd-spatial/Datasets> (acc. August 10, 2012).

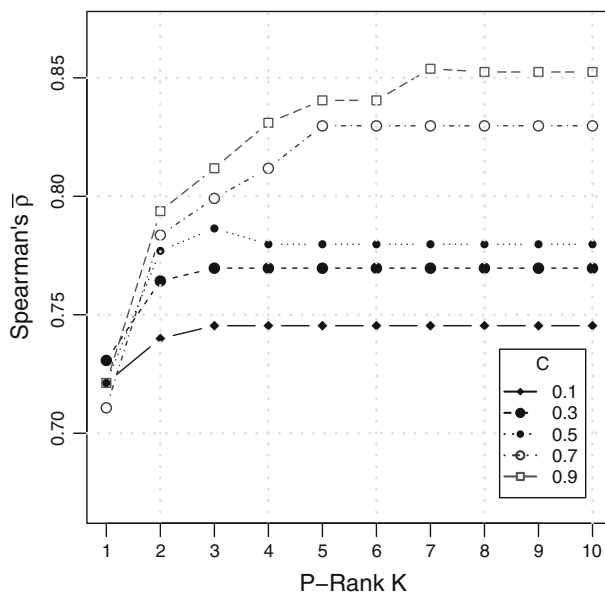


Fig. 2 Experiment results grouped by C (fixed parameter: $\lambda = 1$). K is the P-Rank iteration, while Spearman's $\bar{\rho}$ is a measure of correlation with human behaviour. $p < .0001$ for all $\bar{\rho}$

tion coefficients had to be carried out for each combination of parameters, across the five questions.

Among the existing meta-analytical methods for correlation coefficients, Field [14] concludes that the Hunter–Schmidt method tends to provide the most accurate estimates. This method was originally developed for Pearson's product moment correlation coefficient [24]. As Altman and Gardner [3] noted, both Pearson's r and Spearman's ρ follow a similar statistical distribution, so that the Hunter–Schmidt method can also be applied in our case.

The aggregated $\bar{\rho}$ is computed through a weighted mean, where the weights are the number of pairs in each question. $\bar{\rho}$ expresses the overall correlation between the rankings of P-Rank applied to the OSM Semantic Network, and the MDSM human-generated data set. To assess the statistical significance of these 550 tests, we utilised the Hunter–Schmidt method, based on the standard deviation, the standard error and the Z score [24]. For all of 550 combinations, we obtained $p < .0001$, indicating high statistical significance.

6.3 Discussion of results

The concept rankings for 550 statistically significant cases were generated on the OSM Semantic Network, obtaining correlations with human ranked-pairs of the MDSM data set. Considering only incoming links ($\lambda = 1$), the mean correlation $\bar{\rho}$ is plotted in Fig. 2. A convergence with $K > 7$ can be observed. The similarity scores fluctuate during the first iterations and then plateau, remaining stable in the following iterations. As is reasonable to expect, the convergence is more rapid when C is close to 0. Figure 3 focuses instead on the parameter λ , showing its impact on the correlation. As λ gets closer to 1, the correlation improves steadily, suggesting that incoming links are more relevant to the computation of similarity than outgoing ones.

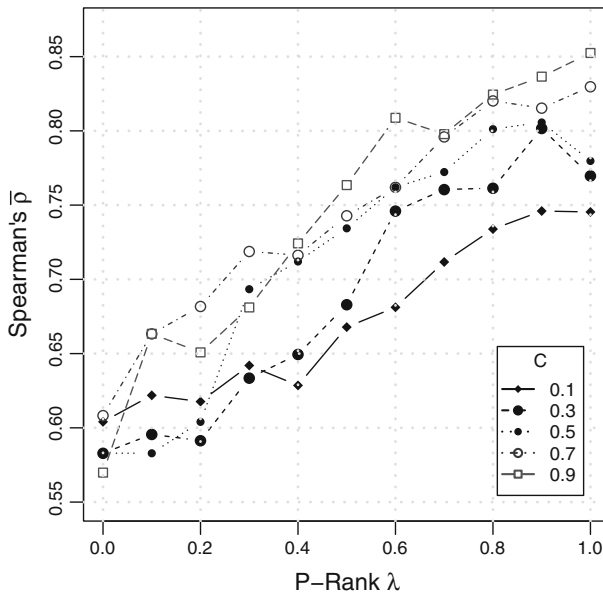


Fig. 3 Experiment results grouped by C (fixed parameter: $K = 10$). K is the P-Rank iteration, while Spearman's $\bar{\rho}$ is a measure of correlation with human behaviour. $p < .0001$ for all $\bar{\rho}$

The overall impact of the decay factor C is clear, as the best correlations are consistently obtained when $C = .9$, and the worst when $C = .1$ (see Figs. 2, 3). Given that $C \in (0, 1)$, it is important to look at its impact at the asymptotes. When $C \rightarrow 0$, the similarity function $s(a, b)$ tends to R_0 . On the other hand, the similarity function does not converge to a finite value when $C \rightarrow 1$. These properties were confirmed on an additional experiment run with $C \in (.9, 1)$, $K = 100$, $\lambda = 1$. With $C \geq .99$, the similarity scores present strong variations even when $K > 50$, not showing any sign of convergence. According to Jeh and Widom [29], the choice of the optimal value of C depends on the specific domain in which SimRank is being applied. On experimental grounds, we can state that, in the context of the OSM Semantic Network, optimal $C \in [.9, .95]$, which suggests that, to match human judgement, similarity has to flow across the graph edges with a slow decay.

The overall results of the experiment are reported in Table 3, which shows the mean Spearman's ρ with 95 % confidence intervals, highlighting the overall cognitive plausibility of the algorithms. Among the non-iterative algorithms ($K = 1$), Small's co-citation performs better than its counterparts. It is possible to notice that, among the iterative algorithms ($K > 1$), SimRank with a low decay ($C = .9$) clearly outperforms the other approaches, reaching a $\bar{\rho} = .85 \pm .07$. Stronger decay factors make the algorithm lose valuable information. The worst results are instead obtained by rvs-SimRank ($\bar{\rho} = .57 \pm .12$), indicating that, in the OSM Semantic Network, outgoing connections between concepts are not strongly correlated with their semantic similarity. This suggests that, when describing concepts in the OSM Wiki, contributors tend to mention similar concepts together.

On the other hand, citations of the same concept while defining similar classes are statistically less common. For example, considering the links between three OSM tags, *waterway=riverbank* references *waterway=river* and *waterway=stream*, two highly similar concepts. The *waterway=river* tag back links *waterway=riverbank*, whilst *waterway=stream*

Table 3 Experimental results

	K	λ	C	Algorithm	Spearman $\bar{\rho}$
K , λ and C are the P-Rank parameters. The Spearman rank correlation is the average of the correlations for each of the five questions of the modified MDSM data set. $\bar{\rho}$ is shown with the 95 % confidence interval computed with the Hunter–Schmidt method [10]. (*) Best performance	1	0	—	Coupling [31]	.55 \pm .09
	1	.5	—	Amsler [4]	.67 \pm .07
	1	1	—	Co-citation [54]	.72 \pm .08
	10	0	.9	rvs-SimRank [62]	.57 \pm .12
	10	0	.5	—	.57 \pm .1
	10	0	.1	—	.60 \pm .07
	10	.5	.9	P-Rank [62]	.76 \pm .08
	10	.5	.5	—	.73 \pm .09
	10	.5	.1	—	.67 \pm .07
	10	1	.9	SimRank [29]	.85 \pm .07*
	10	1	.5	—	.78 \pm .09
	10	1	.1	—	.75 \pm .07

does not. Hence, in this case, incoming links from *waterway=riverbank* strengthen similarity between *waterway=river* and *waterway=stream*, while outgoing links do not encode similarity.

The OSM Semantic Network contains several types of edges (see Sect. 4), which are treated equally in this experiment. In order to assess the importance of each edge type, we ran a series of additional experiments including only one type of edge at a time. The co-citation algorithms are not computable when including only sparse edge types such as *osmwiki:key* and *osmwiki:implies*. On the other hand, when including only edges of type *osmwiki:link*, all the algorithms are computable and the corresponding $\bar{\rho}$ are slightly lower than those obtained in the main experiment with all edge types (e.g. .84 \pm .07 for SimRank, instead of .85 \pm .07). This indicates that the generic hyperlinks *osmwiki:link* convey the bulk of the semantic similarity contained in the network, and the other edges give a minor semantic contribution.

Overall, the results show a clear improvement as λ moves from 0 to 1, and C from 1 to .9. The complete experimental results are available online.²³ The results outlined in this section show that the SimRank algorithm applied to the OSM Semantic Network closely matches the human judgement in the modified MDSM similarity data set, reaching the correlation $\bar{\rho} = .85 \pm .07$ averaged over the five questions.

This can be compared with the MDSM evaluation by Rodríguez and Egenhofer [50]. The MDSM approach was tested on a geographic ontology derived from the combination of definitions in WordNet and in the Spatial Data Transfer Standard (SDTS). This ontology contains formal knowledge carefully encoded by experts, including parts, functions and attributes [13,39]. A comparison between the results of the two approaches is reported in Table 4.

These results indicate that, notwithstanding the lack of rich formal semantics in OSM, it is possible to extract a plausible semantic similarity measure from its crowdsourced semantic network, matching closely the performance obtained on a knowledge-rich formal ontology such as WordNet and SDTS. Based on the collected evidence, we deem that SimRank on the OSM Semantic Network offers a viable tag-to-tag semantic similarity measure for OSM data.

²³ <http://github.com/ucd-spatial/Datasets> (acc. August 10, 2012).

Table 4 Detailed results for SimRank
($C = .9$, $\bar{\rho} = .85 \pm .07$)

Question	Target concept	SimRank ρ (dataset: OSM Semantic Net- work)	MDSM ρ (dataset: Word- Net/SDTS)
<i>QA1</i>	Stadium	.85	.96
<i>QB1</i>	Athletic field	.87	.92
<i>QA4</i>	Travelway	.95	.9
<i>QB4</i>	Path	.9	.88
<i>QAB5</i>	Lake	.7	.82*
–	–	$\bar{\rho} = .85$	$\bar{\rho} = .89$

MDSM results published in [50].
 $\bar{\rho}$ are the weighted means over
the five questions. For all
 ρ , $p < .05$. (*) Mean of survey
A and B

7 Conclusions and future work

In this article, we have presented a contribution to applied knowledge-based systems in the geographic domain, particularly in the area of Volunteered Geographic Information and OpenStreetMap. We have presented (1) the development of the OSM Semantic Network by means of a web crawler tailored on the OSM Wiki website; (2) a study on the cognitive plausibility of co-citation measures to compute semantic similarity of geographic classes in the OSM Semantic Network. Based on the results obtained, the following conclusions can be drawn:

- The OSM Semantic Network²⁴ captures meaningful relationships between geographic concepts in OSM, providing a semantic tool for information retrieval, information integration and data mining. As the OSM Wiki website changes, the crawler enables the regular extraction of an up-to-date graph over time.
- In a semantic network presenting a dense link structure, semantic similarity of concepts can be computed through co-citation algorithms. Such an approach can be successfully applied to compute semantic similarity of geographic classes in the OSM Semantic Network.
- The co-citation algorithms appear cognitively plausible, showing a positive correlation with human judgement. In particular, SimRank obtains the highest plausibility ($\bar{\rho} = .85 \pm .07$) over Small's Co-citation, Amsler, Coupling, rsv-SimRank and P-Rank. This result closely matches the MDSM algorithm applied to WordNet/SDTS classes [50].
- In the context of the OSM Semantic Network, co-citation algorithms consistently obtain a higher plausibility when assuming that concepts are similar when 'they are referenced by similar entities' than when 'they reference similar entities' [62].

The results presented in this paper suggest several research directions. Firstly, the tag-to-tag similarity measure extracted from the OSM Semantic Network can be integrated into a comprehensive OSM similarity framework, enabling an object-to-object metric. An OSM semantic similarity measure should combine network similarity, as well as text similarity and geospatial similarity (geo-location and area). The similarity framework formalised by Janowicz et al. [28] can provide solid theoretical grounds.

²⁴ <http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork> (acc. August 10, 2012).

OSM is far from being the only crowdsourced project modelling general geographic concepts. Notable cases are DBpedia, GeoWordNet and GeoNames²⁵ [7, 17, 41]. Co-citation measures in these knowledge bases can be utilised not only to assess the cognitive plausibility of similarity measures, but also to support information integration [6] and knowledge extraction [23]. As a starting point for future work towards the automatic extension and integration with the Semantic Geospatial Web, our OSM Semantic Network is linked to Wikipedia and LinkedGeoData.

From a cultural perspective, this work is focused on the English parts of the OSM Wiki, introducing a typical Anglo-American bias. One of the key aspects of OSM is the possibility to map local features, which are directly relevant to its contributors, resulting in diverse national and regional communities [34]. In this context, it would be easy to extend the OSM Wiki Crawler to include the numerous non-English pages of the OSM Wiki.²⁶ Co-citation measures are language independent by definition and could be applied to the non-English concepts of OSM in a way analogous to that presented in this paper.

Finally, the results described in Sect. 6.3 highlight a striking difference of cognitive plausibility of co-citation techniques when considering incoming or outgoing links. This indicates that the OSM Wiki contributors have a tendency to cite similar classes together, rather than cite the same class from similar classes. This behaviour might be related to missing links between concepts, which have been detected in Wikipedia [1]. To what degree this phenomenon is generalisable to other contexts is an open question that deserves further investigation, as it would enable a better understanding of how similarity flows across the edges of crowd-sourced semantic networks.

We believe that investigating crowdsourced semantic networks for geographic knowledge will provide valuable contributions to GIScience and, in particular, to the VGI field. In this article, we have described the development of the OSM Semantic Network, evaluating the cognitive plausibility of co-citation measures. This approach has shown a generally high cognitive plausibility, mimicking human rankings of geographic concepts, and can be applied in geographic recommender systems, data mining, location-based services, and in—now unforeseeable—novel neogeographic web applications.

Acknowledgments The research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support. They also wish to thank the anonymous reviewers for their valuable suggestions, and Prof. Leslie Daly (UCD School of Public Health, Physiotherapy & Population Science) for his insightful comments on statistical meta-analysis.

References

1. Adafre S, de Rijke M (2005) Discovering missing links in Wikipedia. In: Proceedings of the 3rd international workshop on link discovery ACM, pp 90–97
2. Agirre E, Alfonseca E, Hall K, Kravalova J, Paşca M, Soroa A (2009) A study on similarity and relatedness using distributional and Wordnet-based approaches. In: Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics ACL, pp 19–27
3. Altman D, Gardner M (1988) Statistics in medicine: calculating confidence intervals for regression and correlation. *Br Med J (Clin Res Ed)* 296(6631):1238–1242
4. Amsler R (1972) Applications of citation-based automatic classification. Technical report 14. Linguistics Research Center, Austin

²⁵ <http://www.geonames.org> (acc. August 10, 2012).

²⁶ http://wiki.openstreetmap.org/wiki/Category:Projects_by_country (acc. August 10, 2012).

5. Auer S, Lehmann J, Hellmann S (2009) LinkedGeoData: adding a spatial dimension to the web of data. In: Proceedings of the international semantic web conference, ISWC 09' vol 5823 of LNCS. Springer, Berlin, pp 731–746
6. Ballatore A, Bertolotto M (2011) Semantically enriching VGI in support of implicit feedback analysis. In: Proceedings of the web and wireless geographical information systems international symposium (W2GIS 2011), vol 6574 of LNCS. Springer, Berlin, pp 78–93
7. Ballatore A, Wilson D, Bertolotto M (2012) A survey of volunteered open geo-knowledge bases in the semantic web. In: Advanced techniques in web intelligence—3: quality-based information retrieval. Studies in computational intelligence, Springer (in press)
8. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web. *Comput Netw* 33(1–6):309–320
9. Collins A, Loftus E (1975) A spreading-activation theory of semantic processing. *Psychol Rev* 82(6): 407–428
10. Diener M, Hilsenroth M, Weinberger J (2009) A primer on meta-analysis of correlation coefficients: the relationship between patient-reported therapeutic alliance and adult attachment style as an illustration. *Psychother Res* 4–5(19):519–526
11. Dolbear C, Hart G (2008) Ontological bridge building—using ontologies to merge spatial datasets. In: Proceedings of the AAAI spring symposium on semantic scientific knowledge integration, AAAI/SSKI'08', AAAI, pp 26–28
12. Egenhofer M (2002) Toward the semantic geospatial web. In: Proceedings of the 10th ACM international symposium on advances in geographic information systems, ACM, pp 1–4
13. Fegeas R, Cascio J, Lazar R (1992) An overview of FIPS 173, the spatial data transfer standard. *Cartogr Geogr Inf Sci* 19(5):278–293
14. Field A (2001) Meta-analysis of correlation coefficients: A monte carlo comparison of fixed- and random-effects methods. *Psychol Methods* 6(2):161–180
15. Formica A, Pourabbas E (2009) Content based similarity of geographic classes organized as partition hierarchies. *Knowl Inf Syst* 20(2):221–241
16. Gartner G, Bennett D, Morita T (2007) Towards ubiquitous cartography. *Cartogr Geogr Inf Sci* 34(4): 247–257
17. Giunchiglia F, Maltese V, Farazi F, Dutta B (2010) GeoWordNet: a resource for geo-spatial applications. In: The semantic web: research and applications, ESWC 2010, vol 6088 of LNCS. Springer, Berlin, pp 121–136
18. Goodchild M (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221
19. Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ Plan B Plan Des* 37(4):682–703
20. Haklay M, Singleton A, Parker C (2008) Web mapping 2.0: the neogeography of the GeoWeb. *Geogr Compass* 2(6):2011–2039
21. Haklay M, Weber P (2008) OpenStreetMap: user-generated street maps. *IEEE Pervasive Comput* 7(4): 12–18
22. Halpin H, Robu V, Shepherd H (2007) The complex dynamics of collaborative tagging. In: Proceedings of the 16th international conference on world wide web, ACM, pp 211–220
23. Hu B (2010) Wiki'mantics: interpreting ontologies with Wikipedia. *Knowl Inf Syst* 25(3):445–472
24. Hunter J, Schmidt F (1990) Methods of meta-analysis: correcting error and bias in research findings. SAGE, Newbury Park
25. Janowicz K, Keßler C, Panov I, Wilkes M, Espeter M, Schwarz M (2008) A study on the cognitive plausibility of SIM-DL similarity rankings for geographic feature types. In: The European Information Society: taking geoinformation science one step further, LINGC. Springer, Berlin, pp 115–134
26. Janowicz K, Keßler C, Schwarz M, Wilkes M, Panov I, Espeter M, Bäumer B (2007) Algorithm, implementation and application of the SIM-DL similarity server. In: GeoSpatial semantics: second international conference, GeoS 2007, vol 4853 of LNCS. Springer, Berlin, pp 128–145
27. Janowicz K, Raubal M, Kuhn W (2011) The semantics of similarity in geographic information retrieval. *J Spat Inf Sci* 2(1):29–57
28. Janowicz K, Raubal M, Schwing A, Kuhn W (2008) Semantic similarity measurement and geospatial applications. *Trans GIS* 12(6):651–659
29. Jeh G, Widom J (2002) SimRank: a measure of structural-context similarity. In: Proceedings of the 8th ACM international conference on knowledge discovery and data mining, SIGKDD, ACM, pp 538–543
30. Keßler C (2011) What is the difference? A cognitive dissimilarity measure for information retrieval result sets. *Knowl Inf Syst* 30(2):319–340
31. Kessler M (1963) Bibliographic coupling between scientific papers. *Am Doc* 14(1):10–25

32. Kuhn W (2005) Geospatial semantics: why, of what, and how?. In: *Journal of Data Semantics III. Special issue on Semantic-based Geographical Information Systems*, vol 3534 of LNCS. Springer, Berlin, pp 1–24
33. Li P, Liu H, Yu J, He J, Du X (2010) Fast single-pair SimRank computation. In: *Proceedings of the SIAM international conference on data mining, SDM2010*. Omnipress, Madison, pp 571–582
34. Lin Y (2011) A qualitative enquiry into OpenStreetMap making. *New Rev Hypermed Multimed* 17(1): 53–71
35. Lin Z, Lyu M, King I (2011) MatchSim: a novel similarity measure based on maximum neighborhood matching. *Knowl Inf Syst* 32(1):1–26
36. Liu J, Chen H, Furuse K, Kitagawa H, Yu JX (2011) On efficient distance-based similarity search. In: *Proceedings of the 11th IEEE international conference on data mining workshops, IIEEE*, pp 1199–1202
37. Lizorkin D, Velikhov P, Grinev M, Turdakov D (2008) Accuracy estimate and optimization techniques for SimRank computation. In: *Proceedings of the VLDB endowment*, vol 1, very large data base endowment, pp 422–433
38. Mika P (2005) Ontologies are us: a unified model of social networks and semantics. In: *Proceedings of the 4th international semantic web conference, ISWC 2005*, vol 3729 of LNCS. Springer, Berlin, pp 522–536
39. Miller G (1995) WordNet: a lexical database for English. *Commun ACM* 38(11):39–41
40. Miller G, Charles W (1991) Contextual correlates of semantic similarity. *Lang Cogn Process* 6(1):1–28
41. Mirizzi R, Ragone A, Di Noia T, Di E (2010) Ranking the linked data: the case of DBpedia. In: *Proceedings of 10th international conference in web engineering, ICWE 2010*, vol 6189 of LNCS. Springer, Berlin, pp 337–354
42. Mooney P, Corcoran P (2012) Characteristics of heavily edited objects in OpenStreetMap. *Future Internet* 4(1):285–305
43. Mooney P, Corcoran P, Winstanley A (2010) Towards quality metrics for OpenStreetMap. In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, ACM*, pp 514–517
44. Mülligann C, Janowicz K, Ye M, Lee W (2011) Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In: *Spatial information theory*, vol 6899 of LNCS. Springer, Berlin, pp 350–370
45. Nakayama K, Hara T, Nishio S (2008) Wikipedia link structure and text mining for semantic relation extraction. Towards a huge scale global web ontology. In: *Proceedings of the workshop on semantic search (SemSearch 2008)*, 5th European semantic web conference (ESWC 2008), vol 334 of CEUR workshop proceedings, pp 59–73
46. Nitzschke J (2012) OpenStreetMap's growth accelerates. Technical report, BeyoNav, Chicago. <http://www.beyonav.com/openstreetmaps-growth-accelerates>
47. Priedhorsky R, Terveen L (2008) The computational Geowiki: what, why, and how. In: *Proceedings of the ACM conference on computer supported cooperative work, CSCW 2008*. ACM, pp 267–276
48. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th international joint conference on artificial intelligence, IJCAI'95*, vol 1. Morgan Kaufmann, pp 448–453
49. Robu V, Halpin H, Shepherd H (2009) Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Trans Web* 3(4):1–34
50. Rodríguez M, Egenhofer M (2004) Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *Int J Geogr Inf Sci* 18(3):229–256
51. Rubenstein H, Goodenough J (1965) Contextual correlates of synonymy. *Commun ACM* 8(10):627–633
52. Schwering A (2008) Approaches to semantic similarity measurement for geo-spatial data: a survey. *Trans GIS* 12(1):5–29
53. Selçuk Candan K, Li W (2001) On similarity measures for multimedia database applications. *Knowl Inf Syst* 3(1):30–51
54. Small H (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. *J Am Soc Inf Sci* 24(4):265–269
55. Sowa J (1991) *Principles of semantic networks: explorations in the representation of knowledge*. Morgan Kaufmann, San Mateo
56. Spearman C (1904) The proof and measurement of association between two things. *Am J Psychol* 15(1):72–101
57. Sui D (2008) The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. *Comput Environ Urban Syst* 32(1):1–5
58. Turdakov D, Velikhov P (2008) Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In: *Proceedings of the SYRCDIS 2008 colloquium on databases and information systems*, vol 355 of CEUR workshop proceedings

59. Turner A (2006) Introduction to neogeography. O'Reilly Media, Sebastopol
60. Wan X (2008) Beyond topical similarity: a structural similarity measure for retrieving highly similar documents. *Knowl Inf Syst* 15(1):55–73
61. Wittgenstein L [2009 (1953)] *Philosophical investigations*, 4th edn. Blackwell, Chichester (trans: Anscombe GEM)
62. Zhao P, Han J, Sun Y (2009) P-Rank: a comprehensive structural similarity measure over information networks. In: *Proceedings of the 18th ACM conference on information and knowledge management, CIKM '09*, ACM, pp 553–562

Author Biographies



Andrea Ballatore received an M.Sc. in Virtual Reality from the University of Turin (Italy) in 2006. He then worked as software engineer in information systems and virtual reality in Italy and Ireland. In 2009, he began a Ph.D. at the School of Computer Science and Informatics, University College Dublin. His research interests include Volunteered Geographic Information, geographic information systems and science, geo-semantics, semantic similarity, and media studies. He is part of the strategic research cluster StratAG, funded by Science Foundation Ireland.



Michela Bertolotto received a Ph.D. in Computer Science from the University of Genova (Italy) in 1998. Subsequently, she worked as a Postdoctoral Research Associate in the National Center for Geographic Information and Analysis (NCGIA) at the University of Maine, USA. Since 2000, she has been a faculty member at the School of Computer Science and Informatics of University College Dublin, Ireland. Her research interests include web-based and wireless GIS, the semantic geo-Web, spatiotemporal data modelling and visualisation.



David C. Wilson is an Associate Professor in the College of Computing and Informatics at UNC Charlotte. Dr. Wilson is an expert in intelligent software systems and a leading name in the field of Case-Based Reasoning. Dr. Wilson's research emphasises the development of intelligent software systems to bridge the gaps between human information needs and the computational resources available to meet them. It involves the coordination of intelligent systems techniques (artificial intelligence, machine learning, etc.) with geographic, multimedia, database, internet and communications systems in order to elicit, enhance, apply and present relevant task-based knowledge.